Evaluation of Information Retrieval and Text Mining Tools on Automatic Named Entity Extraction

Nishant Kumar[1], Jan De Beer[2], Jan Vanthienen[1], and Marie-Francine Moens[2]

[1] Research Center for Management Informatics,
Katholieke Universiteit Leuven, Belgium
{nishant.kumar, jan.vanthienen}@econ.kuleuven.be
[2] Legal Informatics and Information Retrieval group,
Katholieke Universiteit Leuven, Belgium
{jan.debeer, marie-france.moens}@law.kuleuven.be

Abstract. We will report evaluation of Automatic Named Entity Extraction feature of IR tools on Dutch, French, and English text. The aim is to analyze the competency of off-the-shelf information extraction tools in recognizing entity types including person, organization, location, vehicle, time, & currency from unstructured text. Within such an evaluation one can compare the effectiveness of different approaches for identifying named entities.

1 Introduction

Named Entity Extraction, a subfield of information extraction, also known as NE Recognition (NER), is to recognize structured information, such as proper names (person, location and organization), date & time, and numerical values (currency and percentage) from natural language text. It has also been extended to identify other patterns, such as email addresses, and URLs. We test named entity extraction from text in the context of the INFO-NS [1] project.

We report our work on 4 commercial IR tools[2], which utilises NER techniques to identify meaningful entities from unstructured text. Named entity recognition constitutes a basic operation in the structuring of texts. Its automation within the Belgian Federal Police would tremendously aid operational analysts in the coding (schematisation) of criminal cases, sometimes covering hundreds of pages that otherwise would have to be skimmed manually for the discovery of entities of interest.

2 Evaluation Method

A set of 6 police narrative reports (2 from each language, Dutch, French & English) has been used as the test bed. A human experimenter manually identifies all entities of interest. We do this testing to measure the tools on conformity ([1])

[1] Visit AGORA at http://www.belspo.be/belspo/fedra/prog.asp?l=en&COD=AG
[2] Due to contractual obligations the names has not been disclosed.

S. Mehrotra et al. (Eds.): ISI 2006, LNCS 3975, pp. 666–667, 2006.

and qualitative criteria. We used standard measures of evaluation, namely precision, recall, and the F-measure ([2]) to assess the tools. We treat misalignment between extracted entity mentions and a golden standard of manually extracted mentions consistently in favor of the tools. For example, the extracted entity "Congo" is equated with the full entity name " Republic of congo", when the latter is present in the text.

3 Evaluation Results

Tools	Tool-A	Tool-B	Tool-C	Tool-D
Precision	0.944	0.718	0.976	0.959
Recall	0.440	0.145	0.397	0.681
F-measure	0.601	0.241	0.564	0.797

Results show high precision on the most common entity types (persons, organisations, locations), up to 97%. Recall is very poor however, less than 50%.

4 Conclusions and Future Directions

The evaluated tools rely largely on human editable dictionaries and work on keyword matching, which does not give the right contextual result and therefore decreases the relevancy factor. Some IE tools provide human editable and expandable rule sets. A rule is a simple regular expression, but can be extended to incorporate lexical analysis of the source text and the context of entity may be used to trigger recognition of their type. We also mention the problem of ambiguity, resulting in errors, mostly when it comes to determine the entity type (e.g. locations or organisation named after persons).

From these findings we may assume that the use of dictionaries and/or rule sets, both are limited in scope and in their tolerance towards typographical, compositional, and other kinds of observed variations.

Also none of the evaluated tools offers a learning approach to automated entity recognition, whereas the academic community has made much progress in this field ([3]). An equally important line of research is the extraction of entities within noisy texts ([4]).

References

[1] N. Kumar, J. De Beer, J. Vanthienen, and M.-F. Moens, "Multi-criteria evaluation of information retrieval tools," in *Proceedings of the 8th International Conference on Enterprise Information Systems(ICEIS)*, 2006.
[2] C. J. Van Rijsbergen, *Information Retrieval*, 2nd ed. Butterworths London, 1979.
[3] M.-F. Moens, *Information Extraction: Algorithms and Prospects in a Retrieval Context*. Springer-Verlag, 2006.
[4] M. Chau, J. J. Xu, and H. Chen, "Extracting meaningful entities from police narrative reports," in *Proceedings of the International Conference on Intelligence Analysis*, 2005.